Simple Linear Regression Model

i	x_i	y_i
1	1	1
2	3	3
3	4	2

best fitting line $y = \beta_0 + \beta_1 x$?

$$\begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 1 \\ 3 \\ 2 \end{bmatrix}$$ (*see text or class notes*)

X (design matrix) $\quad \boldsymbol{y}$ (observation vector)

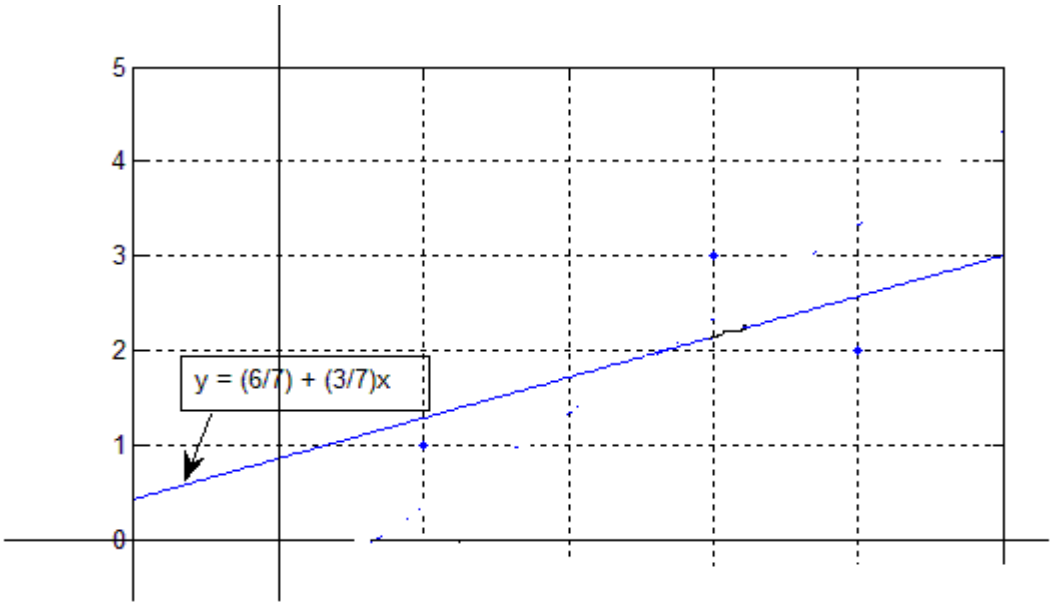
Normal equations $X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$ are

$$\begin{bmatrix} 3 & 8 \\ 8 & 26 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 6 \\ 18 \end{bmatrix}$$

$X^T X$ is invertible (so unique least squares solution)

$$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} = \begin{bmatrix} 3 & 8 \\ 8 & 26 \end{bmatrix}^{-1} \begin{bmatrix} 6 \\ 18 \end{bmatrix} = \begin{bmatrix} \frac{6}{7} \\ \frac{3}{7} \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$$

The regression line <u>for y on x</u>
("for predicting y from x") : $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = \frac{6}{7} + \frac{3}{7}x$



<u>Alternate model</u>:

i	x_i	y_i
1	1	1
2	3	3
3	4	2

best fitting line $x = \alpha_0 + \alpha_1 y$?

$$\begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 2 \end{bmatrix} \boldsymbol{\alpha} = \begin{bmatrix} 1 \\ 3 \\ 4 \end{bmatrix}$$ (*compare to previous matrix equation*)

Y (model matrix) $\boldsymbol{x}$ (observation vector)

Normal Equation $Y^T Y \boldsymbol{\alpha} = Y^T \boldsymbol{x}$

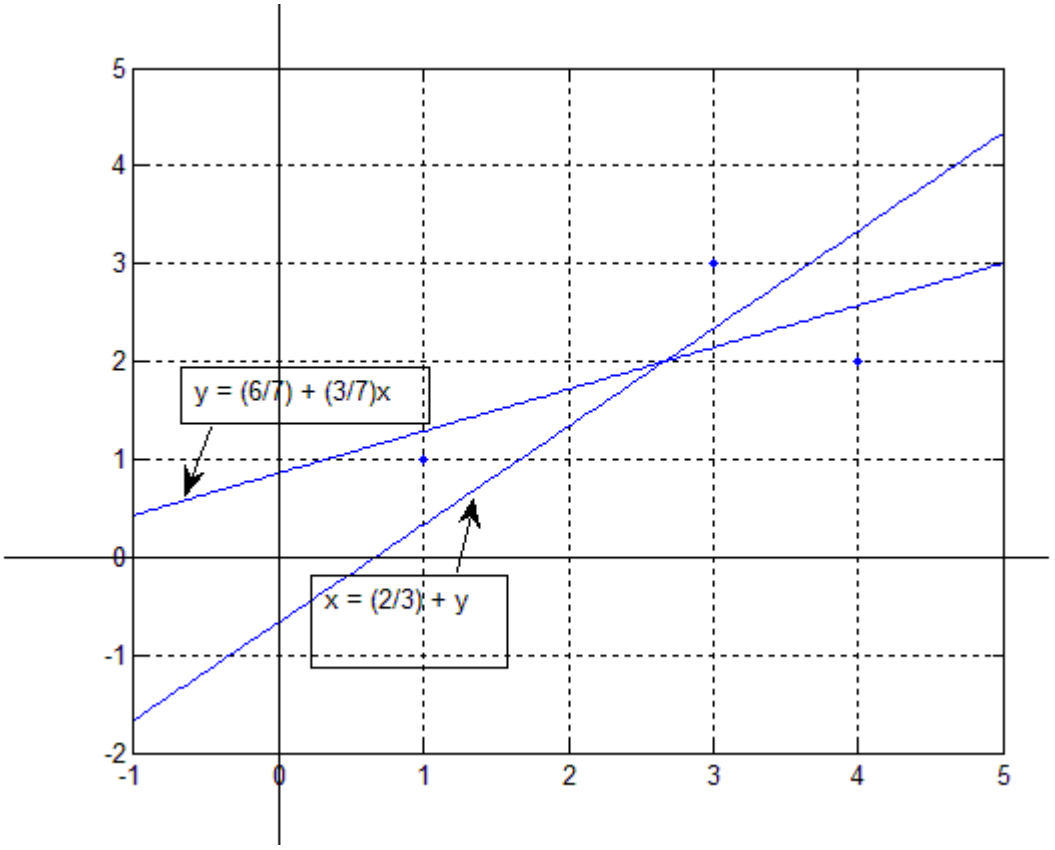
$$\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \boldsymbol{\alpha} = \begin{bmatrix} 8 \\ 18 \end{bmatrix}$$

$$\alpha = \widehat{\alpha} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}^{-1} \begin{bmatrix} 8 \\ 18 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ 1 \end{bmatrix} = \begin{bmatrix} \widehat{\alpha}_0 \\ \widehat{\alpha}_1 \end{bmatrix}$$

The regression line <u>for x on y</u>
("for predicting x from y")

$$\widehat{x} = \widehat{\alpha}_0 + \widehat{\alpha}_1 y = \tfrac{2}{3} + 1y$$

Here are the two regression lines.



Facts: the regression line for x on y is chosen so the the <u>sum of squared horizontal residuals</u> is minimized

the two lines coincide when the data point all line on a straight line (in this case statisticians say that the size of the <u>correlation coefficient</u> for the x, y data is 1 (perfectly linear); the angle between the lines increases when the data points get "further" from lying on a striahgt line is more scattered (unlike a straight line) – the size of the correlation coefficient is clsser to 0.

the two lines always intersect at the poiont $(\overline{x}, \overline{y})$, where $\overline{x}$ is the average of all the x_i's from the data points and $\overline{y}$ is the avaerage of all the y_i's. Usually, $(\overline{x}, \overline{y})$ is not one of the original data points (x_i, y_i).

Example 2 (When the regression line is not unique). We proved in class that there is a unique least squares solution for $X\boldsymbol{\beta} = \boldsymbol{y}$ iff amd and only if X has linearly independent columns. For these regression line examples, it's easy to see that's true unless the data points are <u>all</u> on a vertiocal line.

Simplest possible example: only two data points $(2, 0)$ and $(2, 1)$; they are both on the same vertical line.

$X\boldsymbol{\beta} = \boldsymbol{y}$ is $\begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and the normal equation is

$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$ $\begin{bmatrix} 2 & 4 \\ 4 & 8 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$: infinitely many solutions

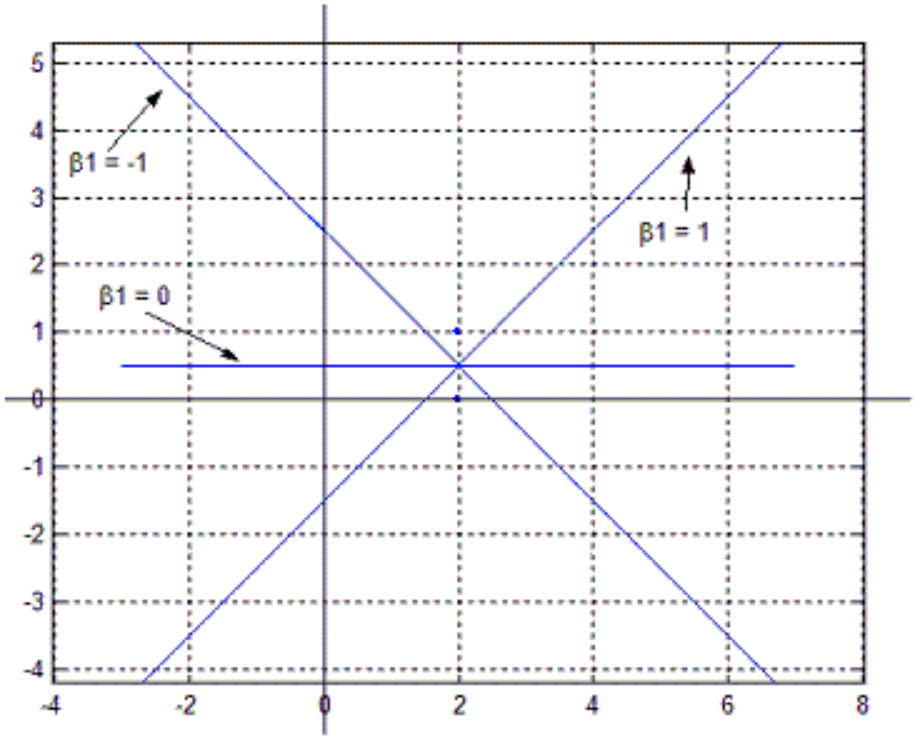
$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ 0 \end{bmatrix} + \widehat{\beta}_1 \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \quad \text{so that}$$

$$\widehat{\boldsymbol{\beta}}_0 = \tfrac{1}{2} - 2\,\widehat{\beta}_1 \text{ (where } \widehat{\beta}_1 \text{ is free)}$$

All of the regression lines pass through the point $(2, \frac{1}{2})$; three of the regression lines

suing $\widehat{\beta}_1 = -1$ $\widehat{\beta}_1 = 0$ $\widehat{\beta}_1 = 1$

are $\frac{5}{2} - 1x$ $\frac{1}{2} + 0x$ $-\frac{3}{2} + 1x$



For <u>each</u> regression line

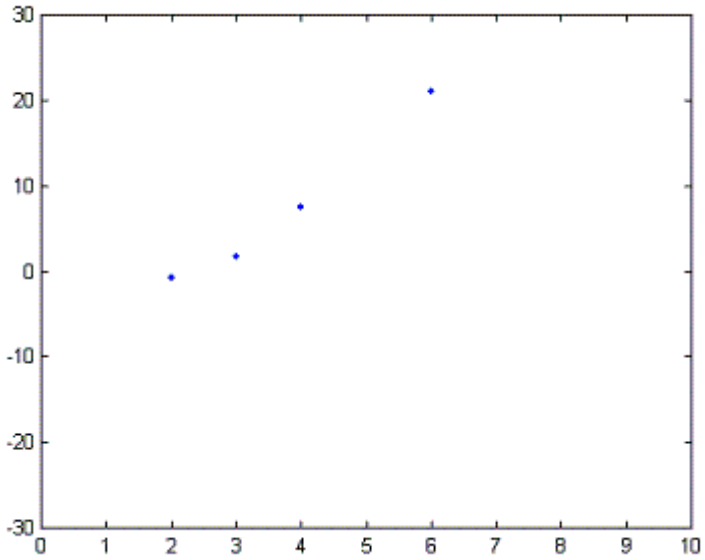
When $x = 2$, the "predicted" observation $\widehat{y} = \frac{1}{2}$

There are only two residuals, for $x = 2$: so each residual error $= y_i - \widehat{y}_i = \frac{1}{2}$, so the

least squares error for <u>any</u> of these line $= \sqrt{(0 - \frac{1}{2})^2 + (1 - \frac{1}{2})^2} = \sqrt{\frac{1}{2}} = \frac{\sqrt{2}}{2} \approx 0.707$

i	x_i	y_i
1	2.1	-0.8
2	3.2	1.6
3	4.5	7.5
4	6.2	21.1
5	7.1	30.1

Find best fit by: $y = \beta_0 + \beta_1 x + \beta_2 x^2$



"Exact fit" would require

$$\begin{aligned} \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 &= y_1 \\ \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 &= y_2 \\ &\vdots \\ \beta_0 + \beta_1 x_5 + \beta_2 x_5^2 &= y_5 \end{aligned} \qquad \text{or}$$

Written in matrix form:

$$\begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_5 & x_5^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_5 \end{bmatrix}$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$
$$X \qquad \boldsymbol{\beta} \quad = \quad \boldsymbol{y}$$

Inconsistent. Find the $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ so that $||\boldsymbol{y} - X\widehat{\boldsymbol{\beta}}||$ is as small as possible.

$X^T X\boldsymbol{\beta} = X^T\boldsymbol{y}$ is (*approximately, after rounding*)

$$\begin{bmatrix} 5 & 23.1 & 123.8 \\ 23.1 & 123.8 & 729.4 \\ 123.8 & 729.4 & 455.32 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 59.5 \\ 381.7 \\ 2493.2 \end{bmatrix}$$

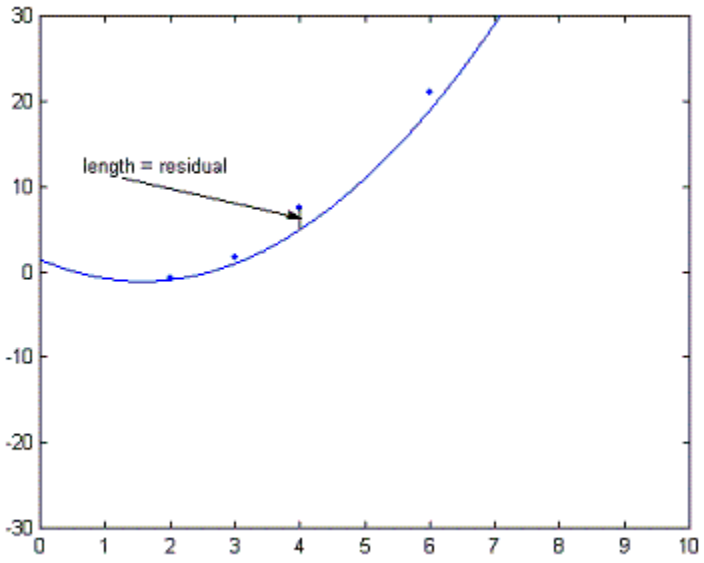
so that (using Matlab)

$$\boldsymbol{\beta} = \widehat{\beta} = \begin{bmatrix} 5 & 23.1 & 123.8 \\ 23.1 & 123.8 & 729.4 \\ 123.8 & 729.4 & 455.32 \end{bmatrix}^{-1} \begin{bmatrix} 59.5 \\ 381.7 \\ 2493.2 \end{bmatrix}$$

$$\approx \begin{bmatrix} 1.4759 \\ -3.2591 \\ 1.0295 \end{bmatrix} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix}.$$

The "best fitting quadratic" is $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$

$$\approx 1.4759 - 3.2591x + 1.0295x^2$$



It is the "best fit" in the sense that this particular quadratic minimizes the sum of the squared residuals

You might think this example is "nonlinear" because a quadratic is involved. But the problem is really about finding the best linear combination of the functions $1, x, x^2$ to approximate the data points: this expresses is itself in a numeric matrix equation to solve.

Find best fitting <u>plane</u> to the data points

$(x_1, y_1, z_1), \ldots, (x_n, y_n, z_n)$ **in** $\mathbb{R}^3$**.**

$$z = \beta_0 + \beta_1 x + \beta_2 y$$

$$\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & y_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{where}$$

$$\uparrow \qquad \uparrow \qquad \uparrow$$
$$X \qquad \boldsymbol{\beta} \qquad \boldsymbol{y}$$

For example: $(1, 2, 3),\ (0, 2, 4),\ (-1, -2, 1)$, and $(1, 1, 0)$. A perfect fit would require:

$$\begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & -1 & -2 \\ 1 & 1 & 1 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 3 \\ 4 \\ 1 \\ 0 \end{bmatrix}$$

The normal equation is

$$\begin{bmatrix} 4 & 1 & 3 \\ 1 & 3 & 5 \\ 3 & 5 & 13 \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} 8 \\ 2 \\ 12 \end{bmatrix}$$

and the least squares solution is

$$\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}} \approx \begin{bmatrix} 1.4783 \\ -2.2174 \\ 1.4348 \end{bmatrix}$$

(*rounded to 4 decimal places*)

The best fitting plane is

$$z = 1.4783 - 2.2174x + 1.4348y$$

(for each x_i, y_i then equation "predicts" a point on the plane $(x_i, y_i, \widehat{z}_i)$, and there is an actual data point (x_i, y_i, z_i). The best fitting plane (from the least squares solution) is the one that minimizes $\sum (z_i - \widehat{z}_i)^2 =$ the sum of the squared residuals $(z_i - \widehat{z}_i)$